**实验一**

1. 当前大数据技术的基础是由\_\_\_\_\_\_首先提出的。（ C）

A：微软

B：百度

C：谷歌

D：阿里巴巴

2. 大数据的起源是\_\_\_\_\_\_。（ C ）。

A：金融

B：电信

C：互联网

D：公共管理

3. 根据不同的业务需求来建立数据模型，抽取最有意义的向量，决定选取哪种方法的数据分析角色人员是\_\_\_\_\_\_。（C）

A：数据管理人员

B：数据分析员

C：研究科学家

D：软件开发工程师

4. 大数据的最显著特征是\_\_\_\_\_\_。（ A）

A：数据规模大

B：数据类型多样

C：数据处理速度快

D：数据价值密度高

5. 当前社会中，最为突出的大数据环境是\_\_\_\_\_\_。（A）

A: 互联网

B：物联网

C：综合国力

D：自然资源

**实验二**

1. \_\_\_\_\_\_不是流行的Linux操作系统。（B）

A. CentOS Linux B. Mac OS C. Ubuntu Linux D. Red Flag Linux

2. Linux操作系统删除文件命令为\_\_\_\_\_\_。(D)

A. mkdir B. rmdir C. mv D. rm

3. 在vi中退出不保存的命令是\_\_\_\_\_\_。(D )

A. :q B. :w C. :wq D. :q!

4. 下列关于大数据的分析理念的说法中，错误的是\_\_\_\_\_\_。（ D ）

A：在数据基础上倾向于全体数据而不是抽样数据

B：在分析方法上更注重相关分析而不是因果分析

C：在分析效果上更追究效率而不是绝对精确

D：在数据规模上强调相对数据而不是绝对数据

5. 下列论据中， 能够支撑“大数据无所不能”的观点的是\_\_\_\_\_\_。（ A）

A：互联网金融打破了传统的观念和行为

B：大数据存在泡沫

C：大数据具有非常高的成本

D：个人隐私泄露与信息安全担忧

**实验三**

1. 改变文件所有者的命令为\_\_\_\_\_\_。( A)

A. chmod B. touch C. chown D cat

2. SELECT语句的完整语法较复杂，但至少包括的部分是\_\_\_\_\_\_。 (B )

A．仅 SELECT

B．SELECT ，FROM

C．SELECT，GROUP

D．SELECT，INTO

3. SQL语句中的条件用以下哪一项来表达\_\_\_\_\_\_。 (C )

A．THEN

B．WHILE

C. WHERE

D．IF

4. 创建数据库使用以下哪项\_\_\_\_\_\_。 ( D)

A．create mytest

B．create table mytest

C．database mytest

D．create database mytest

5.对于Redis，下列说法错误的是\_\_\_\_\_\_。（ D ）

A. Redis是一个开源的使用ANSI C语言编写、支持网络、可基于内存亦可持久化的日志型.Key-Value数据库，并提供多种语言的API。

B. Redis是一个key-value存储系统

C. Redis是一个高性能的key-value数据库。

D. Redis不支持主从同步。

**实验四**

1. Zookeeper的作用是\_\_\_\_\_\_。（ A ）

A. 分布式协调

B. 作业监控

C. 日志收集

D. 集群管理

2.下面关于Zookeeper的特性描述错误的是\_\_\_\_\_\_。（ D ）

A. 客户端所发送的更新会按照他们被发送的顺序进行应用

B. 一条消息要被超过半数的Server接收，他将可以成功写入磁盘

C. 消息更新只能成功或失败，没有中间状态

D. Zookeeper节点数必须为奇数个

3. 查看Kafka的某Topic的partition详细信息时，使用如下哪个命令\_\_\_\_\_\_。（C）

A. bin/kafka-topic.sh –create

B. bin/kafka-topic.sh –list

C. bin/kafka-topic.sh –describe

D. bin/kafka-topic.sh --delete

4. 下列哪些组件必须依赖于Zookeeper才能运行\_\_\_\_\_\_。( A )

1. HDFS HA 2. HBase 3. Spark 4. YARN

A. 1和2 B. 2和3 C. 3和4 D. 1和4

5. HBase依赖\_\_\_\_\_\_提供消息通信机制。( A )

A. Zookeeper

B. Chubby

C. RPC

D. Socket

**实验五**

1. 下列哪个是 Hadoop 运行的模式\_\_\_\_\_\_。（D）

A. 单机版

B. 伪分布式

C. 分布式

D. 以上都正确

2. Hadoop 作者是\_\_\_\_\_\_。（C）

A. Martin Fowler

B. Kent Beck

C. Doug cutting

D. Yoshua Bengio

3. 以下四个Hadoop预定义的Mapper实现类的描述错误的是\_\_\_\_\_\_。（B）

A．IdentityMapper<K, V>实现Mapper<K, V, K, V>，将输入直接映射到输出

B．InverseMapper<K, V>实现Mapper<K, V, K, V>，反转键/值对

C. RegexMapper<K>实现Mapper<K, Text, Text, LongWritable>，为每个常规表达式的匹配项生成一个(match, 1)对

D．TokenCountMapper<K>实现Mapper<K, Text, Text, LongWritable>，当输入的值为分词时，生成(taken, 1)对

4. 关于Hadoop单机模式和伪分布式模式的说法，正确的是\_\_\_\_\_\_。（D）

A．两者都起守护进程，且守护进程运行在一台机器上

B．单机模式不使用HDFS，但加载守护进程

C．两者都不与守护进程交互，避免复杂性

D．后者比前者增加了HDFS输入输出以及可检查内存使用情况

5. 配置Hadoop时，JAVA\_HOME包含在哪一个配置文件中\_\_\_\_\_\_。（B）

A．hadoop-default.xml

B．hadoop-env.sh

C．hadoop-site.xml

D．configuration.xsl

**实验六**

1. 以下Hive sql语法正确的是\_\_\_\_\_\_。（D）

A. select \* from a inner join b on a.id&lt;&gt;b.id

B. select \* from a where a.id in (select id from b)

C. select sum(a.amt) as total from a where a.total&gt;20

D. select \* from a inner join b on a.id=b.id

2. 下面有关HIVE描述错误的是\_\_\_\_\_\_。（D）

A. Hive的集合数据类型有map，struct，array这三种

B. hive.mapred.mode=strict表示所有查询语句都必须指定分区

C. hive.auto.convert.join为mapjoin开关

C. parquet是行式存储，orc是列式存储

3. 下面关于Hive的说法正确的是\_\_\_\_\_\_。（A）

A. Hive是基于Hadoop的一个数据仓库工具，可以将结构化的数据文本映射为一张数据库表，并提供简单的SQL查询功能

B. Hive可以直接使用SQL语句进行相关操作

C. Hive能够在大规模数据集上实现低延迟快速的查询

D. Hive在加载数据过程中不会对数据进行任何的修改，只是将数据移动到HDFS中Hive设定的目录下

4. 有关HIVE中ORDER BY 和 SORT BY 用法正确的是\_\_\_\_\_\_。 (B )

A. SORT BY 用于分组汇总

B. SORT BY 用于局部排序，ORDER BY用于全局排序

C. 使用完全一致

D. 其他说法都不对

5. 在Hive中下列哪个命令不可以实现去重\_\_\_\_\_\_。（D）

A. distinct

B. group by

C. row\_number

D. having

**实验七**

1. Spark相比MapReduce的优势不包括\_\_\_\_\_\_。（B）

A. DAG执行引擎，中间结果不落盘

B. 线程池模型增多task启动开销

C. 充分利用内存，减少磁盘IO

D. 更适合迭代计算

2. 以下属于spark支持的join类型有\_\_\_\_\_\_。（D）

A. inner join

B. left outer join

C. right outer join

D. 以上都支持

3. 以下哪个概念与Spark无关\_\_\_\_\_\_。（C）

A. Graphx

B. Structrued Streaming

C. Spout

D. Data Frames

4. Spark不支持使用哪种语言进行开发\_\_\_\_\_（C）\_。

A. Java

B. C++

C. Scala

D. Python

5. 以下哪个不是spark的组件\_\_\_\_\_\_。（A）

A. Spark R

B. Spark Streaming

C. Mllib

D. GraphX

**实验八**

1. 以下选项中，Python网络爬虫方向的第三方库是\_\_\_\_\_。（A）

A. Scrapy

B. Openpyxl

C. PyQt5

D. Numpy

2. Chrome浏览器如何打开调试模式\_\_\_\_\_。（D）

A. 按F2

B. 按F8

C. 按F10

D. 按F12

3. 以下Scrapy爬虫项目目录和对应功能正确的有\_\_\_\_\_。（D）

A. items.py，定义结构化数据

B. pipelines.py，爬取后的数据的处理逻辑

C. spiders，爬虫程序目录

D. 以上都正确

4. 下面关于Scrapy的命令的功能正确的有\_\_\_\_\_。（D）

A. scrapy startproject，创建项目

B. scrapy genspider，创建爬虫

C. scrapy crawl，运行爬虫

D.以上都正确

5. 以下哪个不是GET请求和POST请求的区别\_\_\_\_\_。（A）

A. get传送的数据量较大，；post传送的数据量较小

B. GET请求没有POST请求安全，重要数据容易暴露

C. GET请求一般用于获取服务器资源；POST请求用于在服务器上创建资源

D. GET请求将参数设置在url中，以？和&得形式拼接；POST请求将参数设置在请求体中

实操题：

### 实验任务一：安装和配置MySQL

#### 步骤一：删除系统自带的MariaDB：

执行以下命令找到系统自带MariaDB并删除：

rpm -qa | grep mariadb



yum remove mariadb-libs-5.5.44-2.el7.centos.x86\_64



查询已安装的MariaDB组件：

rpm -qa | grep mariadb



根据查询结果依次执行以下命令删除查询到的组件：

rpm -e --nodeps （查询到的组件名）



#### 步骤二：安装依赖的组件

执行以下命令安装net-tools组件：

yum install -y net-tools



#### 步骤三：上传MySQL安装包

通过文件上传工具将安装文件[mysql-5.7.18-1.el7.x86\_64.rpm-bundle.tar](https://download.oracle.com/otn-pub/java/jdk/8u201-b09/42970487e3af4f5aa5bca3f542482c60/jdk-8u201-linux-x64.tar.gz)上传到/opt目录



#### 步骤四：解压MySQL安装包

执行以下命令解压MySQL安装包

cd /opt

tar -xvf [mysql-5.7.18-1.el7.x86\_64.rpm-bundle.tar](https://download.oracle.com/otn-pub/java/jdk/8u201-b09/42970487e3af4f5aa5bca3f542482c60/jdk-8u201-linux-x64.tar.gz)

1

#### 步骤五：安装MySQL组件

执行以下命令安装MySQL相关组件：

rpm -ivh mysql-community-common-5.7.18-1.el7.x86\_64.rpm

rpm -ivh mysql-community-libs-5.7.18-1.el7.x86\_64.rpm

rpm -ivh mysql-community-client-5.7.18-1.el7.x86\_64.rpm

rpm -ivh mysql-community-server-5.7.18-1.el7.x86\_64.rpm



#### 步骤六：起停数据库

查看MySQL是否启动：

service mysqld status



启动MySQL：

service mysqld start



停止MySQL：

service mysqld stop



重启MySQL：

service mysqld restart



#### 步骤七：初始化root帐号密码

查看临时密码：

grep password /var/log/mysqld.log



登录数据库：mysql -p ，然后根据提示符输入临时密码



执行命令修改密码：

mysql> set password = password(“Gzdsj\_1234”);



mysql\_upgrade -u root -p gzdsj1234



#### 步骤八：设置允许远程连接

登录MySQL，执行以下命令：

mysql> GRANT ALL PRIVILEGES ON \*.\* TO ‘root’@’%’ IDENTIFIED BY ‘Gzdsj\_1234’ WITH GRANT OPTION;



mysql> flush privileges;

